Анализ профилей в социальных сетях

В.А. Бакаев[a], А.В. Благов[a]

[a] *Самарский национальный исследовательский университет имени академика С.П. Королёва, 443086, Московское шоссе, 34, Самара, Россия*

Аннотация

Статья посвящена анализу данных и связей в различных социальных сетях. Предлагается метод отыскания профилей, принадлежащих одним и тем же пользователям, основанный на анализе имеющихся у профиля связей и сообществ. Создан программный комплекс, реализующий данный метод.

Ключевые слова: data mining; социальные сети; графы; алгоритм Label Propagation; Apache Spark

1. Введение

В настоящее время одним из наиболее актуальных направлений в информационных технологиях является анализ данных или Data Mining. Анализ данных представляет собой процесс обнаружения пригодных к использованию сведений в крупных наборах данных, зачастую разнородных. Обычно такие сведения нельзя обнаружить при традиционном просмотре и переборе, поскольку связи слишком сложны, либо из-за чрезмерного объёма (количества).

В социальных сетях генерируются большие потоки данных (создаются профили, связи, контент). Анализируя эти данные можно получить много полезной информации как по различным группам, сообществам и обсуждениям, так и по каждому пользователю в отдельности [1,2].

Большой интерес к социальным сетям испытывают различные коммерческие организации, использующие их как инструмент взаимодействия с аудиторией. Применяя специализированные сервисы, компании анализируют информацию о пользователях, их активностях и персонализируют предложения для отдельно взятых сегментов своей целевой аудитории, тем самым повышая конверсию и снижая затраты на рекламную кампанию.

В статье предлагается метод повышения эффективности подобного рода инструментов и сервисов, который основан на психологии и паттернах человеческого поведения.

Предлагаемый метод основан на следующих фактах:

- многие пользователи Интернета имеют аккаунты сразу в нескольких популярных социальных сетях (ВКонтакте, Facebook, Instagram и Twitter);
- многие пользователи социальных сетей скрывают информацию о себе от незнакомых людей (в том числе информацию о наличии аккаунтов в других социальных сетях);
- поскольку социальные сети являются предметом социализации людей, то для каждого пользователя можно выделить хотя бы одно сообщество людей такое, что пользователи этого сообщество попарно знакомы друг с другом;
- человек имеет аккаунты в разных социальных сетях и контактирует с одними и теми же людьми.

Для реализации метода разрабатывается программный комплекс, который:

а) анализирует все профили целевых социальных сетей и на основе публичных данных находит аккаунты, принадлежащие владельцу исходного профиля;

б) для пользователей, на страницах которых отсутствует информация о наличии у них аккаунтов в других социальных сетях, находит связи с другими социальными сетями на основании сообществ, в которых состоит пользователь.

2. Сбор первичных данных и скорость их обработки

На первом этапе система анализирует всех пользователей социальных сетей ВКонтакте, Twitter и Instagram и группирует их по следующим правилам:

- в каждой группе находится не более одного профиля из каждой социальной сети;
- все профили внутри одной группы принадлежат одному человеку.

Эта задача решается при помощи программного каркаса Apache Spark (в частности, надстройки Spark Streaming, предназначенной для потоковой обработки данных см. рис. 1) и брокера сообщений RabbitMQ, реализующего доставку исходных данных в Spark Streaming.

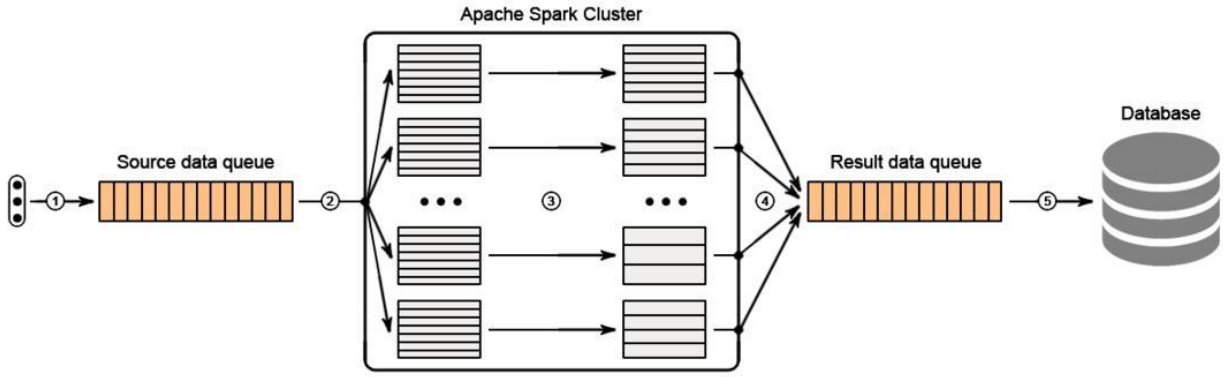

Рис. 1. Архитектура агрегатора пользователей социальных сетей (pipeline).

Описание шагов:

1. Добавление данных из различных источников в очередь для дальнейшей обработки. Данные представляют собой набор пар *(network_id, user_id)*, содержащих информацию о профилях, которые требуется проанализировать.
2. Формирование RDD (Resilient Distributed Dataset) путём пакетирования исходных данных, находящихся в очереди, для увеличения производительности.
3. Преобразование RDD (mapping). Для каждой пары (network_id, user_id) алгоритм находит и группирует профили в других социальных сетях, а также дополнительную информацию о человеке, которому принадлежит исходный аккаунт. Алгоритм запускается повторно для каждого найденного профиля до тех пор, пока не будет найдена вся доступная информация о пользователе. В качестве источников может быть: публичная информация, указанная на странице пользователя (статус, контактная информация, записи в ленте и т.д.).
4. Экспорт полученных данных из RDD в очередь для последующего сохранения.
5. Сохранение результатов в NoSQL базе данных MongoDB в виде документов со структурой, отраженной в таблице 1.

Скорость обработки данных составляет примерно 120-130 профилей в секунду. Для работы использовалась виртуальная машина Microsoft Azure A2 v2 (2 ядра, 4Гб RAM, 20Гб SSD). Анализировались случайные пользователи социальной сети ВКонтакте (1.000.000 профилей).

Таким образом, если допустить, что скорости обработки профилей ВКонтакте, Instagram и Twitter равны, получим приближенную оценку времени, которое потребуется для анализа всех пользователей целевых социальных сетей:

$$T(n) = \frac{4*10^8 + 6*10^8 + 13*10^8}{120n} \approx 5324 \text{ часа } \approx 221 \text{ день},$$

где n - количество серверов в кластере со схожей конфигурацией.

При горизонтальном масштабировании кластера наблюдается линейная зависимость между количеством серверов и скоростью обработки профилей.

Пример ссылки на таблицу: результаты эксперимента отражены в таблице 1.

Таблица 1. Структура хранения данных профилей

Название поля	Тип	Описание
_id	objectId	идентификатор документа в коллекции
vk_id	int32	идентификатор профиля в сети «ВКонтакте»
facebook_id	int64	идентификатор профиля в сети Facebook
instagram_id	int64	идентификатор профиля в сети Instagram
twitter_id	int64	идентификатор профиля в сети Twitter
other	Object	дополнительная информация (номер телефона, адрес электронной почты, skype и т.д.)

Дальнейшая задача сводится к расширению полученной базы путём объединения профилей по описанным ранее правилам, на страницах которых не указаны бэк-линки на другие социальные сети.

3. Предварительная обработка данных

Основываясь на гипотезе, что человек состоит в одних и тех же сообществах во всех социальных сетях, которыми пользуется, мы можем установить связь между профилями разных социальных сетей, находящихся в одном сообществе и с определённой вероятностью предположить, что они принадлежат одному человеку.

Целью данного этапа является выделение сообществ среди людей, находящихся в базе данных, полученной из предыдущего шага. Для этого используется расширение Apache Spark GraphX, которое предназначено для распределенной обработки графов.

Алгоритм предварительной обработки данных:

1. Генерация графа на основе имеющихся данных. Вершины представляют собой сущность “человек” и хранят идентификаторы профилей, принадлежащих конкретному пользователю. Связь между вершинами устанавливается по следующему принципу: две вершины смежны, если связны профили соответствующих социальных сетей. Определим вес ребра между вершинами как:

$$w(A,B) = \left|\{i : i \in [0, n-1] \cap \exists A_i, B_i \cap rel(A_i, B_i)\}\right|, \quad (1)$$

где *rel(a, b)* - функция, которая истина тогда и только тогда, когда профили *a* и *b* взаимосвязаны (установлено отношение дружбы или проявление активности).

2. К полученному графу применяется алгоритм Label Propagation [3], реализованый внутри GraphX API, который решает задачу кластеризации и находит сообщества в графе.
3. Для каждого сообщества генерируется RDD с набором профилей ВКонтакте, Facebook, Instagram и Twitter, которые связаны с одним или несколькими членами исходного сообщества.

Таким образом формируется набор данных (dataset), на основе которого мы можем строить предположения о принадлежности группы аккаунтов разных социальных сетей (без явного указания взаимосвязей) одному человеку.

4. Группирование данных на основе анализа общих черт

Данный этап является заключительным в решении поставленной задачи. К каждому набору данных из датасета применяется следующая процедура:

1. Построение полного многодольного графа, в котором хранится информация о профилях социальных сетей и потенциал, характеризующий вероятность их принадлежности одному человеку;

   В вершинах графа содержится информация о профилях, которая используется при их сравнении.

   Для сравнения двух профилей используется многослойная нейронная сеть. На входной слой сети подаётся вектор размерности 12, содержащий следующие данные:

- Name ↔ Name'
- max(Name → Username', Name' → Username)
- max(Name → E-mail', Name' → E-mail)
- max(Name → Skype', Name' → Skype)
- Username ↔ Username'
- max(Username → E-mail', Username' → E-mail)
- Username ↔ Skype'
- max(Skype → Username', Skype' → Username)
- max(Skype → E-mail', Skype' → E-mail)
- E-mail ↔ E-mail'
- Phone ↔ Phone'
- Website ↔ Website'

Определим операции $a \to b$ и $a \leftrightarrow b$:

1.1 Полнота вхождения a в b:

$$a \to b = 1 - \frac{d + r + s}{len(a)} \in [0, 1],$$

где *d* - количество операций удаления для преобразования *a* в *b*;

r - количество операций замены для преобразования *a* в *b*;

s - количество операций транспозиции для преобразования *a* в *b*;

len(x) - функция вычисления длины аргумента.

1.2 Сравнение *a* и *b*:

$$\forall i \in [1, len(a)], j \in [1, len(b)] d[i,j] = 1 - \frac{dist(a[i], b[j])}{len(b[j])} \in [0, 1],$$

$$a \leftrightarrow b = \frac{\sum_1^{len(a)} d[i, fit(i)]}{min(len(a), len(b))} \in [0, 1],$$

где *dist(a, b)* - функция, вычисляющая расстояние Дамерау-Левенштейна [4] для строк *a* и *b*;

fit(i) - функция, возвращающая индекс слова строки b, поставленного в соответствие слову *a[i]*.

Операция сравнения не учитывает порядок слов. Все слова исходных строк попарно сравниваются, а затем, при помощи алгоритма Куна-Манкреса [5], каждому слову строки a ставится в соответствие слово строки b так, чтобы сумма схожести по всем парам слов была максимальной. Также не учитываются знаки препинания и прочие символы (за исключением букв и цифр).

Перед обработкой все символы входных данных приводятся к латинским по правилам транслитерации. Для этого используется сводная таблица основных алфавитов (русский, украинский, болгарский, индийский, арабский).

Обучающая и контрольная выборки собраны на основе первичных данных. Размер обучающей выборки ~106 пар.

В качестве отрицательных примеров использовались как случайные пары профилей, так и пары, найденные при помощи полнотекстового поиска по разным параметрам (*name, username, email, skype*).

2. В сгенерированном графе для каждой пары долей выполняется следующий алгоритм:
   2.1. рёбра сортируются в порядке убывания весов;
   2.2. удаляются рёбра, вес которых меньше порогового значения или одна из инцидентных вершин уже связана с какой-либо вершиной противоположной доли.

В результате этих преобразований получается граф, в котором каждая компонента связности представляет собой группу аккаунтов из разных социальных сетей, которые принадлежат одному человеку.

В связи с тем, что человек может одновременно принадлежать нескольким сообщества, а также если одна и та же группа была сформирована в нескольких сообществах, то можно полагать, что аккаунты этой группы действительно принадлежат одному пользователю.

Данные, полученные на последнем этапе, записываются в ту же базу, где хранятся первичные данные. Однако они не используются в качестве входных данных для реализованного алгоритма ввиду своей недостоверности.

5. Заключение

Обработка и анализ данных социальных сетей позволяет персонализировать продукт или услугу для конкретного сегмента целевой аудитории. Полученный в результате работы программный комплекс стирает границы между социальными сетями для различных сервисов, позволяя им объединять API и оперировать сущностью “человек”, а не “профиль”, что делает их работу эффективнее.

Дальнейшее исследование может быть продолжено в части модернизации алгоритмов агрегирования и анализа полученных данных. Это необходимо для реализации решения следующих задач:

- выявление популярных социальных сетей и сервисов, которые пользователи упоминают на своих страницах (например, LinkedIn, Last.fm) и разработка парсеров для них;
- анализ дополнительных источников информации на страницах пользователей в ВКонтакте, Instagram и Twitter (например, лента новостей в Twitter);
- анализ страниц пользователей в целевых социальных сетях и поиск дополнительных параметров, на основании которых можно сравнивать профили на принадлежность одному пользователю.

Литература

[1] Tan, W. Social-network-sourced big data analytics/ W. Tan, M.B. Blake, I. Saleh, S. Dustdar //IEEE Internet Computing, 2013. №. 5 – P. 62-69.

[2] Khotilin, M.I. Visualization and Cluster Analysis of Social Networks / M.I. Khotilin , A.V. Blagov // CEUR Workshop Proceedings, 2016. – Vol.1638.– P.843-850.

[3] Liu, X. Advanced modularity-specialized label propagation algorithm for detecting communities in networks // X. Liu, T.Murata // Physica A: Statistical Mechanics and its Applications, Vol. 389, Issue 7, P. 1493-1500.

[4] Нечёткий поиск в тексте и словаре. Сметанин Н. [Электронный ресурс]. – Режим доступа: https://habrahabr.ru/post/114997.

[5] Венгерский алгоритм решения задачи о назначениях [Электронный ресурс]. – Режим доступа: http://e-maxx.ru/algo/assignment_hungary.

[6] Apache Spark Documentation [Электронный ресурс]]. – Режим доступа: http://spark.apache.org/docs/2.1.0.